



Image

Docket No.: PC-0025 CIP


Response Under 37 C.F.R. 1.116 - Expedited Procedure

Examining Group 1646

Certificate of Mailing

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to:

Mail Stop: Appeal Brief-Patents, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on March 19, 2004

By:  Printed: Jeannie G. Labra

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE**

In re Application of: Michael G. Walker

Title: ANKYRIN REPEAT DOMAIN 2 PROTEIN VARIANT

Serial No.: 09/758,593

Filing Date: January 10, 2001

Examiner: Li, R.

Group Art Unit: 1646

Mail Stop: Appeal Brief-Patents

Commissioner for Patents

P.O. Box 1450

Alexandria, VA 22313-1450

**TRANSMITTAL FEE SHEET**

Sir:

Transmitted herewith are the following for the above-identified application:

1. Return Receipt Postcard;
2. Brief on Appeal, including Appendix (pp., in triplicate); and
3. One (1) Reference Attachment (in triplicate).

The fee has been calculated as shown below.

☒ Fee for filing a Brief in support of an Appeal under 37 CFR 1.17(c): \$ 330.00


☒ Please charge Deposit Account No. 09-0108 in the amount of : \$ 330.00

The Commissioner is hereby authorized to charge any additional fees required under 37 CFR 1.16 and 1.17, or credit overpayment to Deposit Account No. 09-0108. **A duplicate copy of this sheet is enclosed.**

Respectfully submitted,

INCYTE CORPORATION

Date: March 19, 2004



David G. Streeter, Ph.D.

Reg. No. 43,168

Direct Dial Telephone: (650) 845-5741

Customer No.: 27904

3160 Porter Drive

Palo Alto, California 94304

Phone: (650) 855-0555

Fax: (650) 845-4166



Docket No.: PC-0025 CIP

**Response Under 37 C.F.R. 1.116 - Expedited Procedure**  
**Examining Group 1646**

Certificate of Mailing

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to:  
Mail Stop Appeal Brief-Patents, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on March 19, 2004  
By: [Signature] Printed: Jeannie G. Labra

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE  
BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES**

In re Application of: Michael G. Walker

Title: ANKYRIN REPEAT DOMAIN 2 PROTEIN VARIANT

Serial No.: 09/758,593

Filing Date: January 10, 2001

Examiner: Li, R.

Group Art Unit: 1646

---

Mail Stop Appeal Brief-Patents  
Commissioner for Patents  
P.O. Box 1450  
Alexandria, VA 22313-1450

**BRIEF ON APPEAL**

Sir:

Further to the Notice of Appeal filed January 22, 2004, and received by the USPTO on January 26, 2004, herewith are three copies of Appellants' Brief on Appeal. Authorized fees include the \$ 330.00 fee for the filing of this Brief.

This is an appeal from the decision of the Examiner finally rejecting claims 1-12 of the above-identified application.

**(1) REAL PARTY IN INTEREST**

The above-identified application is assigned of record to Incyte Genomics, Inc., (now Incyte Corporation ) (Reel 011909, Frame 0532) which is the real party in interest herein.

**(2) RELATED APPEALS AND INTERFERENCES**

Appellants, their legal representative and the assignee are not aware of any related

appeals or interferences which will directly affect or be directly affected by or have a bearing on the Board's decision in the instant appeal.

(3) STATUS OF THE CLAIMS

Claims rejected:	Claims 1-12.
Claims allowed:	(none).
Claims canceled:	Claims 13-20.
Claims withdrawn:	(none).
Claims on Appeal:	Claims 1-12 (A copy of the claims on appeal, as amended, can be found in the attached Appendix).

(4) STATUS OF AMENDMENTS AFTER FINAL

An amendment after Final Rejection was filed in accordance with the Examiner's rejections and suggestions in order to place the application in better condition for allowance or appeal. For purposes of appeal, the proposed amendment will be entered. See Advisory Action, mailed March 4, 2004.

(5) SUMMARY OF THE INVENTION

Appellants' invention is directed to polynucleotides encoding an Ankyrin Repeat Domain 2 protein variant (Ankrd2V:SEQ ID NO:1) and a naturally occurring variant of the amino acid sequence of SEQ ID NO:1 having at least 90% sequence identity to the amino acid sequence of SEQ ID NO:1. The invention is further directed to the polynucleotide sequence of SEQ ID NO:2 encoding SEQ ID NO:1, a specific fragment of SEQ ID NO:2 selected from SEQ ID NOs:4-6, and a variant of SEQ ID NO:2 selected from SEQ ID NOs:7-10. Nucleic acids encoding the Ankrd2V of the present invention were first identified as a cardiac muscle-associated gene through the exhibition of a strong association, or coexpression, with known genes that are specifically cardiac muscle-associated. See specification, at page 9, line 33 through page 10, line 14. Northern analysis further showed the expression of the transcript encoding Ankrd2V in musculoskeletal tissues, with a particularly high abundance in a patient with clear cell sarcoma (MUSCNOT10). See specification, at page 10, lines 23-29.

Ankrd2V is identified as skeletal muscle protein by a high level of sequence identity (88%) with two ankyrin-repeat mouse proteins, Mus musculus Ankrd2 (g9501360), and M. musculus skeletal muscle and cardiac protein (SMCP; g5420272), SEQ ID NOs:11-12, respectively, and by the conservation in Ankrd2V of various structural and functional motifs common to the known proteins. See specification, at page 11, lines 5-14. The claimed polynucleotides are therefore asserted to be useful in the diagnosis and treatment of muscle disorders such as muscle hypertrophy and clear cell sarcoma.

**(6) ISSUES**

1. Whether claims 1-12 directed to Ankrd2V encoding polynucleotide sequences, and fragments and variants thereof meet the enablement requirements of 35 U.S.C. §112, first paragraph. In particular, whether the claimed variants and fragments of SEQ ID NO:2 and variants of SEQ ID NO:1 are sufficiently enabled that the skilled artisan would know how to make and/or use them.

2. Whether claims 1 and 3-12 are in accordance with the written description requirements of 35 U.S.C. §112, first paragraph. In particular, whether an isolated cDNA comprising a nucleic acid sequence encoding a naturally occurring variant of SEQ ID NO:1 having at least 90% identity to the amino acid sequence of SEQ ID NO:1 is sufficiently described that one skilled in the art would recognize applicants possession of said variants at the time the application was filed

**(7) GROUPING OF THE CLAIMS**

**As to Issue 1**

All of the claims are grouped together.

**As to Issue 2**

Claims 1 and 3-12 are grouped together. Claim 2 stands alone

**(8) APPELLANTS' ARGUMENTS**

Claims 1-12 stand rejected under 35 U.S.C. § 112, first paragraph, because the specification does not reasonably provide enablement for one skilled in the art to use the invention commensurate with the scope of the claims. The rejection alleges in particular that:

- while the claimed fragments of SEQ ID NO:2 may be used as a probe for the diagnosis of clear cell sarcoma, the claimed variants may not be used as a probe for the diagnosis of disease conditions associated with differential expression of Ankrd2v because determining the specificity of hybridization is empirical by nature and the effect of mismatches is unpredictable, as taught by Wallace et al. (1987) and Sambrook et al. (1989).
- The specification also fails to provide sufficient information on how to produce naturally occurring variants of SEQ ID NO:1 having at least 90% identity to SEQ ID NO:1. There is no sufficient guidance or working examples on how to make and use the variants in producing transgenic cell lines or organisms which model human disorders. Regarding the sequences recited in claim 2, the specification is silent on how to use these variants of a small portion of SEQ ID NO:2 to produce transgenic cell lines or organisms which model a human disorder, in particular, human clear cell sarcoma. Thus, one skilled in the art would not be able to practice the invention without undue experimentation.
- In addition, claim 6 also remains rejected because of its recitation of “A vector”, which was suggested to be replaced by “An expression vector” in claim 4 to overcome this part of the rejection.

**The rejection of claims 1-12 under 35 U.S.C. § 112, first paragraph for lack of enablement is improper, as the inventions of those claims are sufficiently enabled through the described utilities that one skilled in the art would know how to make and/or use them**

Claim 4 has been amended to recite “An expression vector” as the Examiner has suggested. Applicants disagree that one skilled in the art would not know how to use the claimed polynucleotides as probes, for example, for identifying naturally occurring molecules encoding Ankrd2V, allelic variants, or related molecules (specification, at p. 13, lines 22-26); in arrays to monitor large number of genes simultaneously and to identify genetic variants and mutations (specification, at p. 14, lines 16-24); and for chromosomal mapping (specification, at p. 14, lines 25-29). In particular, applicants disagree that the art, as evidenced by Wallace et al. and

Sambrook et al., cited by the Examiner, support an allegation that the effect of mismatches on the specificity of hybridization of polynucleotide sequences, such as those claimed in the instant application, are “unpredictable” to the extent that the skilled artisan would not know how to use them in the hybridization assays described in the specification without undue experimentation. Indeed the Wallace and Sambrook references cited by the Examiner are both laboratory manuals describing procedures that are now routine in the art.

On the contrary, the overall teachings of Sambrook et al (1989) in the reference cited by the Examiner, support applicants position as described in the specification at page 13, line 22 through page 145, line 15, that the specificity of hybridization is highly predictable based the various conditions that may be modified to control the specificity of hybridization. Such procedures are routine in the art and do not constitute “undue experimentation” The single page in Sambrook cited by the Examiner merely discusses the effect of mismatches in base pairing on one parameter of hybridization specificity, i.e., melting temperature ( $T_m$ ). While the reference does recite that the effect of these mismatches on  $T_m$  can be determined empirically and refers to procedures in the manual for determining these effects, the procedures are routine in the art. The Wallace article (1987) specifically discusses issues associated with the specificity of oligonucleotide probes based on parameters such as oligonucleotide length and G + C content, and discusses methods similar to those described in Sambrook for determining the effects of these parameters on hybridization specificity, again using methods that are routine in the art.

Applicants therefore submit that the skilled artisan would know how to use the claimed polynucleotides in hybridization assays to distinguish between Ankrd2V encoding polynucleotides and related molecules in a sample, in arrays to monitor large number of genes simultaneously and to identify genetic variants and mutations, and for chromosomal mapping without undue experimentation using procedures that are routine in the art for regulating the degree of specificity in the hybridization assays. Withdrawal of the rejection of claims 1-12 under 35 U.S.C. § 112, first paragraph is therefore requested.

Claims 1 and 3-12 stand rejected under 35 U.S.C. § 112, first paragraph as containing subject matter which is not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention. The rejection alleges in particular that:

- claim 1(b) recites an isolated cDNA comprising a nucleic acid sequence encoding a naturally occurring variant of SEQ ID NO:1 having at least 90% identity to the amino acid sequence of SEQ ID NO:1, and there is no disclosure of such variants. The Examiner further stated that the claim is drawn to a genus of nucleic acids that is defined only by partial sequence identity to SEQ ID NO:1, not by complete chemical structure as Applicants argued. The claims do not require that the nucleic acids possess any particular biological activity nor any particular conserved structure, or disclosed distinguishing feature. The specification further fails to provide representative examples of such variants and methods of making such variants.
- Accordingly, the Examiner stated, in the absence of sufficient recitation of distinguishing identifying characteristics, the specification does not provide adequate written description of the claimed genus. The Examiner then cited various case law in support of his position, including *Vas-Cath Inc v. Mahurkar*, *Fiers v. Revel*, and *Amgen v. Chugai Pharmaceutical Co.*

**The rejection of claims 1 and 3-12 under 35 U.S.C. § 112, first paragraph is improper, as the inventions of those claims are sufficiently described in terms of chemical and physical structure that one skilled in the art would recognize applicants possession of them at the time the application was filed.**

Applicants have argued that the requirements necessary to fulfill the written description requirement of 35 U.S.C. 112, first paragraph, are well established by case law as well as by the Patent and Trademark Office's own "Guidelines for Examination of Patent Applications Under the 35 U.S.C. Sec. 112, para. 1", published January 5, 2001. Applicants simply disagree with the Examiner's interpretation of these requirements. For example, *Vas-Cath Inc v. Mahurkar* states:

. . . the applicant must also convey with reasonable clarity to those skilled in the art that, as of the filing date sought, he or she was in possession *of the invention*. The invention is, for purposes of the "written description" inquiry, *whatever is now claimed*. *Vas-Cath, Inc. v. Mahurkar*, 19 USPQ2d 1111, 1117 (Fed. Cir. 1991)

The Patent and Trademark Office's own "Guidelines for Examination of Patent Applications Under the 35 U.S.C. Sec. 112, para. 1", published January 5, 2001, provide that :

An applicant may also show that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics which provide evidence that applicant was in possession of the claimed invention, i.e., **complete or partial structure, other physical and/or chemical properties**, functional characteristics when coupled with a known or disclosed correlation between function and structure, **or some combination of such characteristics. What is conventional or well known to one of ordinary skill in the art need not be disclosed in detail.** If a skilled artisan would have understood the inventor to be in possession of the claimed invention at the time of filing, even if every nuance of the claims is not explicitly described in the specification, then the adequate description requirement is met (emphasis added).

Thus, the written description standard is fulfilled by both what is specifically disclosed and what is conventional or well known to one skilled in the art.

SEQ ID NO:1 and SEQ ID NO:2 are specifically disclosed in the application (see, for example, page 3, lines 24-26). Variants of SEQ ID NO:1 are described, for example, at page 4, lines 22-23, where, in particular, a variant having at least 85% amino acid sequence similarity to SEQ ID NO:1 is described. Incyte clones in which the nucleic acids encoding the human Ankr2V were first identified and libraries from which those clones were isolated are described, for example, at page 10, lines 18-22 of the Specification. Chemical and structural features of SEQ ID NO:1 are described, for example, on page 10, line 30 through page 11, line 14. Given SEQ ID NO:1, and the various chemical and structural features described for SEQ ID NO:1, one of ordinary skill in the art would recognize naturally-occurring variants of SEQ ID NO:1 having 90% sequence identity to SEQ ID NO:1. Accordingly, the Specification provides an adequate written description of the recited polypeptide sequences.

**A. The Specification provides an adequate written description of the claimed "variants" of SEQ ID NO:1.**

The Office Action has further asserted that the claims are not supported by an adequate written description because:

The claims do not require that the nucleic acids possess any particular biological activity (claims 1), nor any particular conserved structure, or other disclosed distinguishing feature (claim 1(b)). Thus the claims are drawn to a genus of nucleic acids that is defined only by partial sequence identity.

(page 9 of the Office Action mailed 5/20/2003)



Such a position is believed to present a misapplication of the law.

**1. The present claims specifically define the claimed genus through the recitation of chemical structure**

Court cases in which "DNA claims" have been at issue commonly emphasize that the recitation of structural features or chemical or physical properties are important factors to consider in a written description analysis of such claims. For example, in *Fiers v. Revel*, 25 USPQ2d 1601, 1606 (Fed. Cir. 1993), the court stated that:

If a conception of a DNA requires a precise definition, such as by structure, formula, chemical name or physical properties, as we have held, then a description also requires that degree of specificity.

In a number of instances in which claims to DNA have been found invalid, the courts have noted that the claims attempted to define the claimed DNA in terms of functional characteristics without any reference to structural features. As set forth by the court in *University of California v. Eli Lilly and Co.*, 43 USPQ2d 1398, 1406 (Fed. Cir. 1997):

In claims to genetic material, however, a generic statement such as "vertebrate insulin cDNA" or "mammalian insulin cDNA," without more, is not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function.

Thus, the mere recitation of functional characteristics of a DNA, without the definition of structural features, has been a common basis by which courts have found invalid claims to DNA. For example, in *Lilly*, 43 USPQ2d at 1407, the court found invalid for violation of the written description requirement the following claim of U.S. Patent No. 4,652,525:

1. A recombinant plasmid replicable in procaryotic host containing within its nucleotide sequence a subsequence having the structure of the reverse transcript of an mRNA of a vertebrate, which mRNA encodes insulin.

In *Fiers*, 25 USPQ2d at 1603, the parties were in an interference involving the following count:

A DNA which consists essentially of a DNA which codes for a human fibroblast interferon-beta polypeptide.

Party Revel in the *Fiers* case argued that its foreign priority application contained an adequate written description of the DNA of the count because that application mentioned a potential method for isolating the DNA. The Revel priority application, however, did not have a description of any particular DNA structure corresponding to the DNA of the count. The court therefore found that the Revel priority application lacked an adequate written description of the subject matter of the count.

Thus, in *Lilly* and *Fiers*, nucleic acids were defined on the basis of functional characteristics and were found not to comply with the written description requirement of 35 U.S.C. § 112; *i.e.*, "an mRNA of a vertebrate, which mRNA encodes insulin" in *Lilly*, and "DNA which codes for a human fibroblast interferon-beta polypeptide" in *Fiers*. In contrast to the situation in *Lilly* and *Fiers*, the claims at issue in the present application define polynucleotides or polypeptides in terms of chemical structure, rather than on functional characteristics. For example, the "variant language" of independent claim 1 recites chemical structure to define the claimed genus:

1. An isolated cDNA ... comprising a nucleic acid sequence encoding a protein selected from the group consisting of:...b) a naturally-occurring variant of the amino acid sequence of SEQ ID NO:1 having at least 90% identity to the sequence of SEQ ID NO:1...

From the above it should be apparent that the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1. In the present case, there is no reliance merely on a description of functional characteristics of the polypeptides recited by the claims. In fact, there is no recitation of functional characteristics. Moreover, if such functional recitations were included, it would add to the structural characterization of the recited polypeptides. The polypeptides defined in the claims of the present application recite structural features, and cases such as *Lilly* and *Fiers* stress that the recitation of structure is an important factor to consider in a written description analysis of claims of this type. By failing to base its written description inquiry "on whatever is now claimed," the Office Action failed to provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in *Lilly* and *Fiers*.

2. **The present claims do not define a genus which is highly variant**

Furthermore, the claims at issue do not describe a genus which could be characterized as highly variant. Available evidence illustrates that the claimed genus is of narrow scope.

In support of this assertion, the Examiner's attention is directed to the reference cited at page 29 of the specification by Brenner et al. ("Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships," Proc. Natl. Acad. Sci. USA (1998) 95:6073-6078). Through exhaustive analysis of a data set of proteins with known structural and functional relationships and with <90% overall sequence identity, Brenner et al. have determined that 30% identity is a reliable threshold for establishing evolutionary homology between two sequences aligned over at least 150 residues. (Brenner et al., pages 6073 and 6076.) Furthermore, local identity is particularly important in this case for assessing the significance of the alignments, as Brenner et al. further report that  $\geq 40\%$  identity over at least 70 residues is reliable in signifying homology between proteins. (Brenner et al., page 6076.)

The present application is directed, *inter alia*, to ankyrin repeat domain 2 proteins related to the amino acid sequence of SEQ ID NO:1. In accordance with Brenner et al, naturally occurring molecules may exist which could be characterized as ankyrin repeat domain 2 proteins and which have as little as 40% identity over at least 70 residues to SEQ ID NO:1. The "variant language" of the present claims recites, for example, polynucleotides encoding "a naturally-occurring amino acid sequence having at least 90% sequence identity to the sequence of SEQ ID NO:1" (note that SEQ ID NO:1 has 329 amino acid residues). This variation is far less than that of all potential ankyrin repeat domain 2 proteins related to SEQ ID NO:1, i.e., those ankyrin repeat domain 2 proteins having as little as 40% identity over at least 70 residues to SEQ ID NO:1.

**3. The state of the art at the time of the present invention is further advanced than at the time of the *Lilly* and *Fiers* applications**

In the *Lilly* case, claims of U.S. Patent No. 4,652,525 were found invalid for failing to comply with the written description requirement of 35 U.S.C. §112. The '525 patent claimed the benefit of priority of two applications, Application Serial No. 801,343 filed May 27, 1977, and Application Serial No. 805,023 filed June 9, 1977. In the *Fiers* case, party Revel claimed the benefit of priority of an Israeli application filed on November 21, 1979. Thus, the written description inquiry in those case was based on the state of the art at essentially at the "dark ages" of recombinant DNA technology.

The present application has a priority date of April 26, 1999. Much has happened in the development of recombinant DNA technology in the 20 or more years from the time of filing of the applications involved in *Lilly* and *Fiers* and the present application. For example, the technique of polymerase chain reaction (PCR) was invented. Highly efficient cloning and DNA sequencing technology has been developed. Large databases of protein and nucleotide sequences have been compiled. Much of the raw material of the human and other genomes has been sequenced. With these remarkable advances one of skill in the art would recognize that, given the sequence information of SEQ ID NO:1, and the additional extensive detail provided by the subject application, the present inventors were in possession of the claimed polynucleotide variants encoding SEQ ID NO:1 at the time of filing of this application.

#### **4. Summary**

The Office Action failed to base its written description inquiry "on whatever is now claimed." Consequently, the Action did not provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in cases such as *Lilly* and *Fiers*. In particular, the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1. The courts have stressed that structural features are important factors to consider in a written description analysis of claims to nucleic acids and proteins. In addition, the genus of polypeptides defined by claim 1 is adequately described, as evidenced by Brenner et al and consideration of the claims of the '740 patent involved in *Lilly*. Furthermore, there have been remarkable advances in the state of the art since the *Lilly* and *Fiers* cases, and these advances were given no consideration whatsoever in the position set forth by the Office Action.

Applicants submit that one skilled in the art would clearly recognize applicants possession of the claimed variants of SEQ ID NO:1 based on the extensive chemical and structural features described for SEQ ID NO:1, and withdrawal of the rejection of claims 1 and 3-12 under 35 U.S.C. § 112, first paragraph for lack of adequate written description is therefore requested.

Due to the urgency of this matter, including its economic and public health implications, an expedited review of this appeal is earnestly solicited.

If the USPTO determines that any additional fees are due, the Commissioner is hereby authorized to charge Deposit Account No. **09-0108**.

**This brief is enclosed in triplicate**

Respectfully submitted,

INCYTE CORPORATION

Date: March 19, 2000 David G. Streeter

David G. Streeter, Ph.D.

Reg. No. 43,168

Direct Dial Telephone: (650) 845-5741

**Customer No.: 27904**

3160 Porter Drive

Palo Alto, California 94304

Phone: (650) 855-0555

Fax: (650) 849-8886

**Enclosures:**

1. Brenner et al., Proc. Natl. Acad. Sci. 95:6073-78 (1998)

**APPENDIX - CLAIMS ON APPEAL**

1. An isolated cDNA, or the complement thereof, comprising a nucleic acid sequence encoding a protein selected from the group consisting of:
  - a) an amino acid sequence of SEQ ID NO:1; and
  - b) a naturally occurring variant of the amino acid sequence of SEQ ID NO:1 having at least 90% identity to the amino acid sequence of SEQ ID NO:1.
2. An isolated cDNA comprising a nucleic acid sequence selected from:
  - a) SEQ ID NO:2 or the complement thereof;
  - b) a fragment of SEQ ID NO:2 selected from SEQ ID NOs 4-6 or the complements thereof; and
  - c) a variant of SEQ ID NO:2 selected from SEQ ID NOs:7-10.
3. A composition comprising the cDNA or the complement of the cDNA of claim 1 and a labeling moiety.
4. An expression vector comprising the cDNA of claim 1.
5. An isolated host cell comprising the vector of claim 4.
6. A method for using a cDNA to produce a protein, the method comprising:
  - a) culturing the host cell of claim 5 under conditions for protein expression; and
  - b) recovering the protein from the host cell culture.
7. A method for using a cDNA to detect expression of a nucleic acid in a sample comprising:
  - a) hybridizing the composition of claim 3 to nucleic acids of the sample, thereby forming hybridization complexes; and
  - b) comparing hybridization complex formation with a standard, wherein the comparison indicates expression of the cDNA in the sample.
8. The method of claim 7 further comprising amplifying the nucleic acids of the sample prior to hybridization.
9. The method of claim 7 wherein the composition is attached to a substrate.
10. The method of claim 7 wherein increased expression of the cDNA in the sample when compared with a standard of normal tissue is diagnostic of clear cell sarcoma.
11. A method of using a cDNA to screen a plurality of molecules or compounds for a molecule or compound that specifically binds the cDNA, the method comprising:
  - a) combining the cDNA of claim 1 with a plurality of molecules or compounds under conditions to allow specific binding; and

b) detecting specific binding, thereby identifying a molecule or compound which specifically binds the cDNA.

12. The method of claim 11 wherein the molecules or compounds are selected from DNA molecules, RNA molecules, artificial chromosome constructions, peptides, transcription factors, and repressors.

## Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER<sup>\*†‡</sup>, CYRUS CHOTHIA<sup>\*</sup>, AND TIM J. P. HUBBARD<sup>§</sup>

<sup>\*</sup>MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and <sup>§</sup>Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

**ABSTRACT** Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA  $ktup = 1$ , and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests have evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

**Previous Assessments of Sequence Comparison.** Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith-Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed ( $ktup = 2$ ) or greater effectiveness ( $ktup = 1$ ). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1998 by The National Academy of Sciences 0027-8424/98/956073-6\$2.00/0  
PNAS is available online at <http://www.pnas.org>.

Abbreviation: EPQ, errors per query.

<sup>†</sup>Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126

<sup>‡</sup>To whom reprints requests should be addressed. e-mail: [brenner@hyper.stanford.edu](mailto:brenner@hyper.stanford.edu).



superfamilies. Pearson found that modern matrices and "ln-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith-Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

**A Database for Testing Homology Detection.** Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it

is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or  $\approx 0.5\%$  of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from <http://sss.stanford.edu/sss/>, and databases derived from the current version of SCOP may be found at <http://scop.mrc-lmb.cam.ac.uk/scop/>.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

**Assessment Data and Procedure.** Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith-Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties  $-12/-1$  (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

**The "Coverage Vs. Error" Plot.** To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

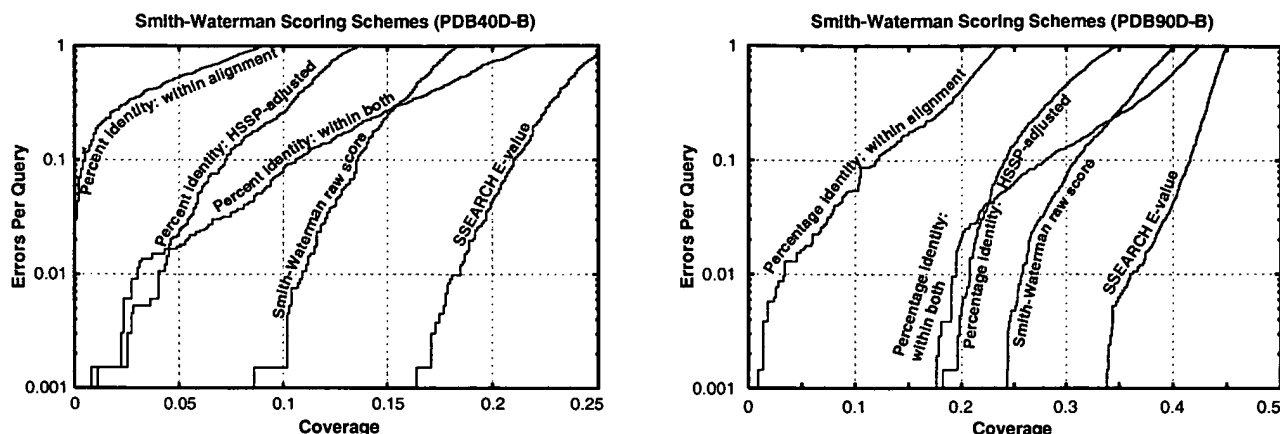


FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith-Waterman. (A) Analysis of PDB40D-B database. (B) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the x axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The y axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The y axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSSP equation (17) is  $H = 290.15l^{-0.562}$  where  $l$  is length for  $10 < l < 80$ ;  $H > 100$  for  $l < 10$ ;  $H = 24.7$  for  $l > 80$ . The percentage identity HSSP-adjusted score is the percent identity within the alignment minus  $H$ . Smith-Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how

protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Receiver Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

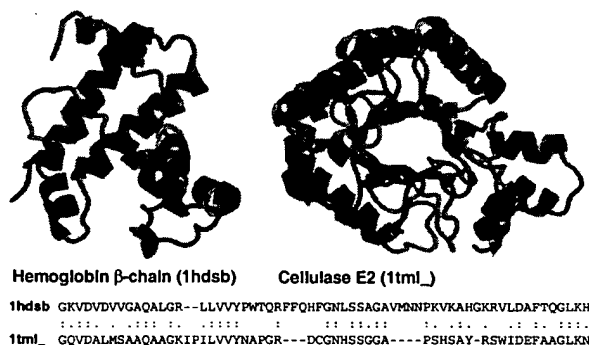


FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin  $\beta$ -chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1tml, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

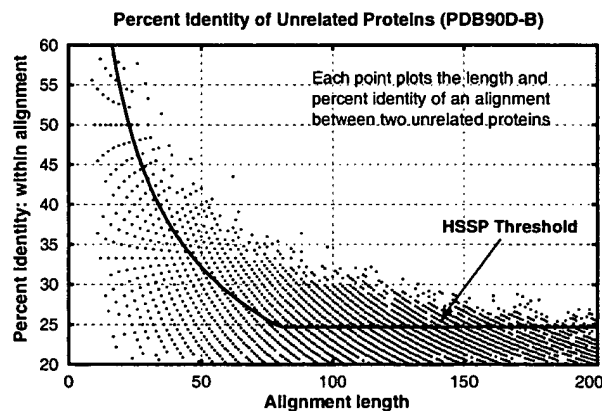


FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

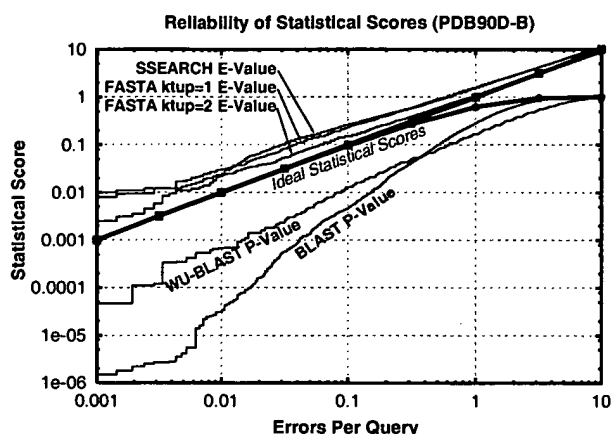


FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

**The Performance of Scoring Schemes.** All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith-Waterman" score, which is the measure optimized by the Smith-Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure

related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

**Sequence Identity.** Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

**Raw Scores.** Smith-Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

**Statistical Scores.** Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

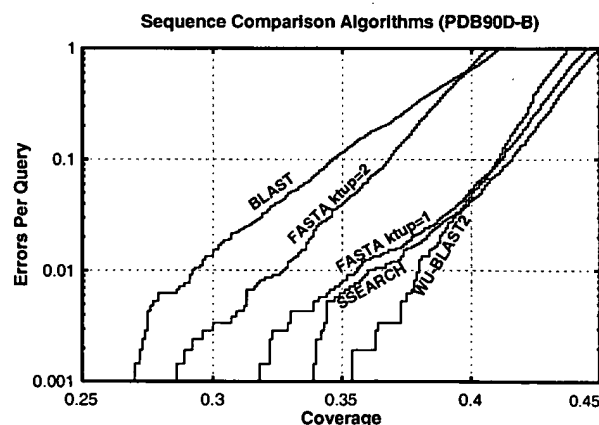
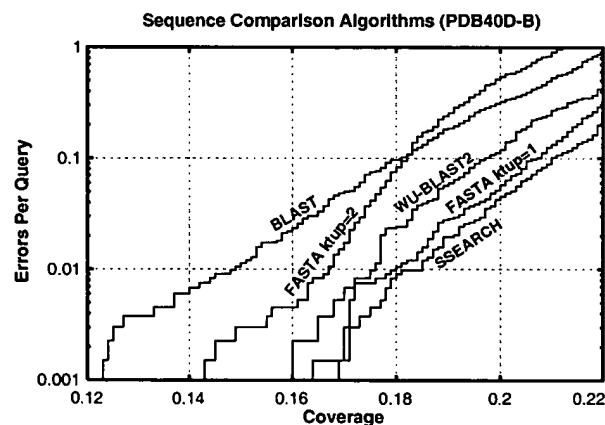


FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (A) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (B) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

**Overall Detection of Homologs and Comparison of Algorithms.** The results in Fig. 5A and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup = 1 is nearly as effective as SSEARCH. FASTA ktup = 2 and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup = 1. WU-BLAST2 is slightly faster than FASTA ktup = 2, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5B). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA ktup = 1, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity

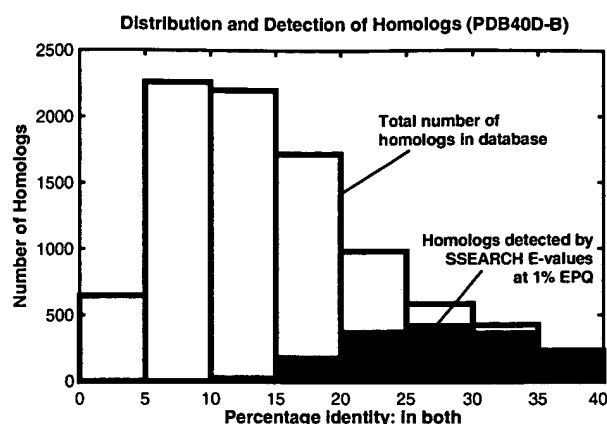


Fig. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pairwise sequence comparison to detect them.

are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

## CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (i) using a large current database in which the protein sequences have been complexity masked and (ii) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith-Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA ktup = 1 E-values	3.9	0.03	17.9
FASTA ktup = 2 E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

\*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.\*\*

\*\*Additional and updated information about this work, including supplementary figures, may be found at <http://sss.stanford.edu/sss/>.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
2. Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266**, 460–480.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247**, 536–540.
5. Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266**, 635–643.
6. Pearson, W. R. (1991) *Genomics* **11**, 635–650.
7. Pearson, W. R. (1995) *Protein Sci.* **4**, 1145–1160.
8. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197.
9. George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* **266**, 41–59.
10. Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249**, 816–831.
11. Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17**, 49–61.
12. Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24**, 21–25.
13. Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24**, 189–196.
14. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919.
15. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Bio-medical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
16. Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
17. Sander, C. & Schneider, R. (1991) *Proteins* **9**, 56–68.
18. Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233**, 716–738.
19. Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1**, 89–94.
20. Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1**, 77–78.
21. Arratia, R., Gordon, L. & M, W. (1986) *Ann. Stat.* **14**, 971–993.
22. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
23. Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 5873–5877.
24. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6**, 119–129.
25. Pearson, W. R. (1996) *Methods Enzymol.* **266**, 227–258.
26. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12**, 215–226.
27. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266**, 554–571.
28. Waterman, M. S. & Vingron, M. (1994) *Stat. Science* **9**, 367–381.
29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13**, 669–678.
30. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
31. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* **7**, 369–376.
32. Orengo, C., Michie, A., Jones S, Jones D. T, Swindells M. B. & Thornton, J. (1997) *Structure (London)* **5**, 1093–1108.
33. Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39**, 561–577.
34. Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20**, 25–33.
35. Fitch, W. M. (1966) *J. Mol. Biol.* **16**, 9–16.
36. Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* **4**, 1123–1127.
37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402.
38. Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131**, 417–433.
39. Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32**, 9906–9916.
40. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20**, 374–376.